Improving Demographic Information for Address Based Sampling (ABS) frames

American Association for Public Opinion Research Annual Conference
New Orleans, LA, May 21st, 2017

Joe McMichael & Jamie Ridenhour

http://artnc.org/works-of-art/new-orleans-ragging-home
What are most doing now?

- Appending area-level demographics
  - Decennial Census
  - American Community Survey

- Appending address-level demographics
  - Vendors (i.e. MSG)

What if we used predictive models built from previous survey?
RTI Enhanced ABS Frame
Enhanced ABS Frame

- CDS (ABS foundation)
- Geocode addresses
- Area-level demographics
  - Census PDB
  - Decennial Census
  - ACS
- Address & person-level
  - Acxiom InfoBase
  - Many sources (black-box)
  - Completeness varies
  - Accuracy varies
- RTI modeled demographics
- City-style, PO Box, etc
- Vacancy status
- Single vs. Multi-family
- Census block group demo
- Child age group
- Adult name
- Adult age
- Adult race
- Income
- Education
- Subscribe to Cat Fancy (joking)
- many others
Methods
Survey Data

Survey A
- Mail survey
- Single state (a large one)
- Sample - 73,000
- Respondents - 26,500
- 36% return rate
- Collected over four years

Survey B
- Multi-mode national survey
- Consider mail screener only
- Sample – 18,000
- Respondents - 8,000
- 44% return rate
Inspired by the Census Return Rate Challenge

- Kaggle competition
  - Census Return Rate Challenge
  - Predict mail return rates for the Decennial Census
  - Top models employed machine learning ensemble methods

- Census used most important predictors
Model building

Random forest for variable selection
- Python 3, SciKit, RandomForestClassifier
- 205 predictors
- Variable importance (see next slide for example)

Logistic Regression
- Top predictors from random forest
- Split into training and validation data
- Stepwise variable selection
- ASE of model with validation data
- SAS HPLOGISTIC
### Variable Importance Example

#### Highest Ranked Variables

- LRS (PDB)
- DOB age 65+ (Acxiom)
- High rise (CDS)
- Vacant (CDS)
- % white alone (Decennial Census)
- PersonicX generation categories (Acxiom)
- % College Grad (ACS)
- Has a surname (Acxiom)
- Has DOB (Acxiom)
- Hispanic Surname (Acxiom)
- Has child (Acxiom)
- DOB 40-49 (Acxiom)
- DOB 50-59 (Acxiom)
- Black Surname (Acxiom)
- % Other Language (ACS)
Area under the curve (AUC)
- Measures classification error of predicted vs. observed
- Range 0.5 to 1
- 1 is perfect
- 0.5 is useless

Density Strata
- 6 evenly sized strata
- Based on the modeled predicted probability
<table>
<thead>
<tr>
<th>Density Stratum</th>
<th>Child in HH 13-17</th>
<th>Hispanic</th>
<th>Education HS or less</th>
<th>Tobacco Use</th>
<th>Adult 18-24</th>
<th>Adult in HH 18-24</th>
</tr>
</thead>
<tbody>
<tr>
<td>High - 1</td>
<td>34%</td>
<td>56%</td>
<td>45%</td>
<td>29%</td>
<td>20%</td>
<td>35%</td>
</tr>
<tr>
<td>2</td>
<td>18%</td>
<td>16%</td>
<td>29%</td>
<td>25%</td>
<td>9%</td>
<td>24%</td>
</tr>
<tr>
<td>3</td>
<td>14%</td>
<td>7%</td>
<td>25%</td>
<td>18%</td>
<td>3%</td>
<td>20%</td>
</tr>
<tr>
<td>4</td>
<td>9%</td>
<td>4%</td>
<td>17%</td>
<td>20%</td>
<td>3%</td>
<td>16%</td>
</tr>
<tr>
<td>5</td>
<td>3%</td>
<td>3%</td>
<td>13%</td>
<td>14%</td>
<td>1%</td>
<td>12%</td>
</tr>
<tr>
<td>Low - 6</td>
<td>1%</td>
<td>2%</td>
<td>10%</td>
<td>11%</td>
<td>1%</td>
<td>8%</td>
</tr>
<tr>
<td>AUC</td>
<td>0.77</td>
<td>0.87</td>
<td>0.69</td>
<td>0.61</td>
<td>0.80</td>
<td>0.67</td>
</tr>
<tr>
<td>Expected Population Prevalence</td>
<td>13%</td>
<td>15%</td>
<td>40%</td>
<td>18%</td>
<td>12%</td>
<td>15%</td>
</tr>
<tr>
<td>Observed Average</td>
<td>13%</td>
<td>15%</td>
<td>23%</td>
<td>19%</td>
<td>6%</td>
<td>19%</td>
</tr>
</tbody>
</table>
Discussion

- Modeled demographics can improve density stratification

- Sample design
  - Useful for rare populations
  - Optimal sample allocations need accurate predicted prevalence

- Data collection interventions and protocols

- Weight adjustments
Next steps

- Feature Engineering
- Explore other ensemble methods
- Evaluate additional survey data and outcomes
- Evaluate additional auxiliary data sources